Statistische Entscheidungs Theorie

Dozent: Dr. Zoran Nikolić

Machine Learning Seminar

Wadim Koslow
Mathematisches Institut

Gliederung

- Machine Learning Einführung
- K-Nächste-Nachbarn
- Least Squares
- Vergleich der Modelle
- Fazit

Machine Learning

- Ziel: Finde Zusammenhang zwischen Input Daten X und Output Daten Y um möglichst genaue Vorhersagen zu treffen:
  - Wir suchen: $$\hat{Y} = \hat{f}(X) \approx f(X) = Y$$

- Input Daten $X \in \mathbb{R}^{N \times K}$ (bzw. Zufallsvariable) werden auch ***Features*** oder ***unabhängige Variablen*** genannt
- $N \mathrel{\widehat{=}}$ #***Beobachtungen***
- $K \mathrel{\widehat{=}}$ #***Merkmale***
- Output Daten $Y \in \mathbb{R}^{N}$ (bzw. Zufallsvariable) werden auch ***Targets*** oder ***abhängige Variablen*** genannt

Vorhersagen

- ***Quantitative*** Vorhersagen: Output Daten Y liegen in einem stetigen Wertebereich
  - Bsp.: Vorhersage von Aktienkurswerten oder Temperatur
  - Bsp.: $Y(X) = 2X + 5$
  - Wird auch als ***Regression*** bezeichnet

Vorhersagen

- ***Qualitative*** Vorhersagen: Output Daten Y bestehen aus Klassen
  - Bsp.: Vorhersage ob Aktienkurs steigt oder sinkt oder ob Patient stirbt oder überlebt
  - Für 2 Klassen: $Y(X) = \begin{cases} 1 & falls\ X \in Klasse\ 1 \\ 0 & falls\ X \in Klasse\ 0 \end{cases}$
  - Für K > 2 Klassen : $Y \in \mathbb{R}^{N \times K}$ mit K binären Vektoren
  - Wird auch als ***Klassifikation*** bezeichnet

Vorhersagen

- Genaue Vorhersagen ⇔ Kleiner Fehler zwischen vorhergesagten und echten Werten $Y - \hat{Y}$

- ***Expected Prediction Error*** *(EPE):*

$$EPE(f) = E\left(\left(Y - f(X)\right)^2\right)$$

- Es gilt: Minimum der Funktion $EPE(f)$ liegt bei

$$f(X) = E(Y|X)$$

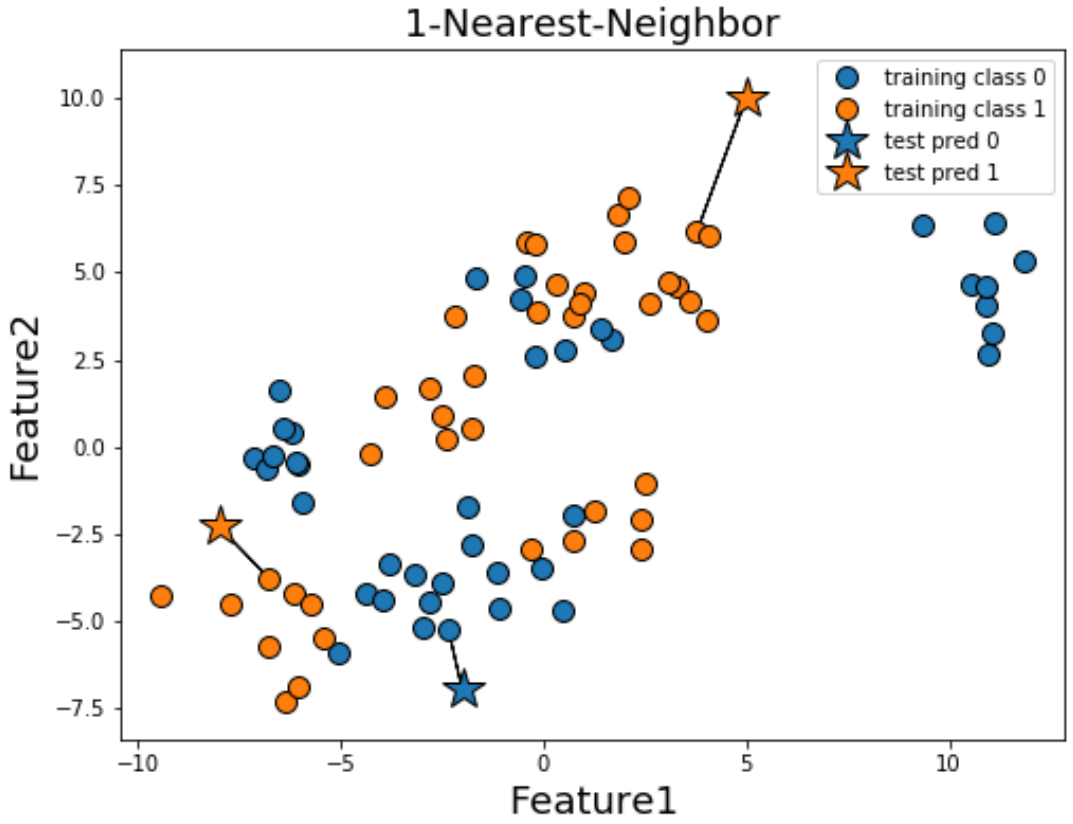
K-Nächste-Nachbarn

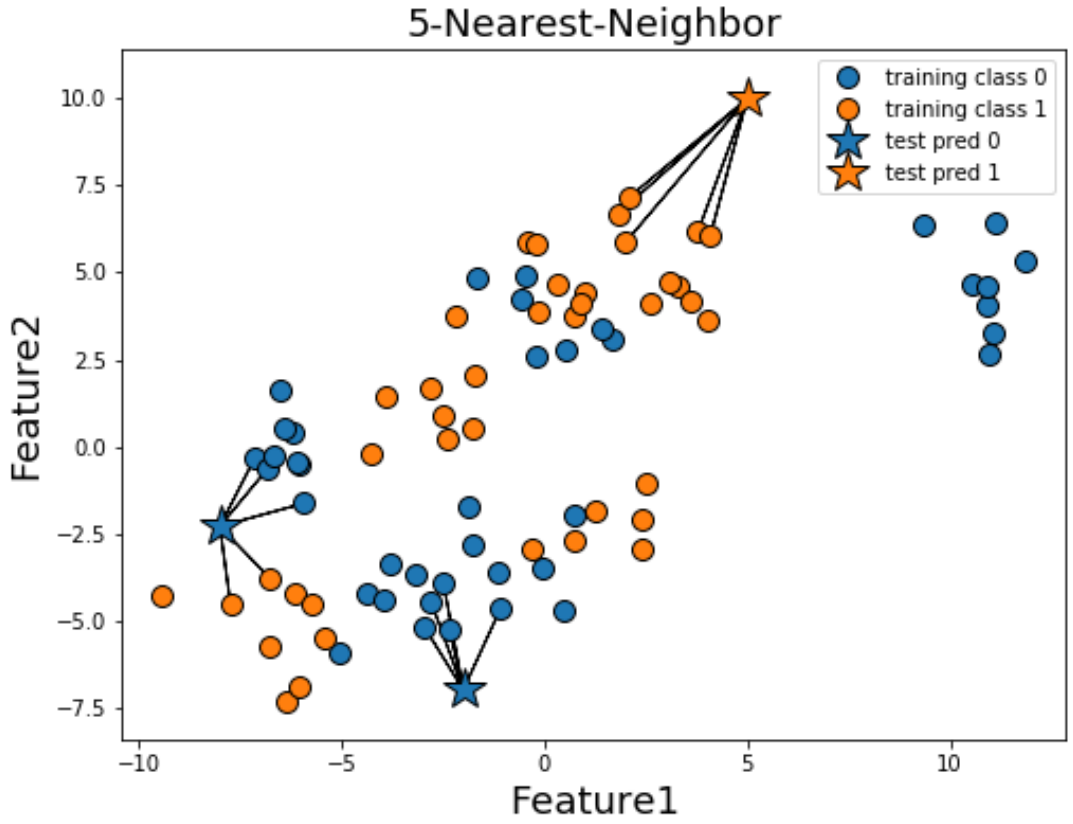
- Ersetze Erwartungswert durch lokalen Mittelwert:

$$f(X) = E(Y|X) \approx \frac{1}{k} \sum_{x_i \in N_k(x)} y(x_i) =: \hat{f}(x)$$

$N_k(x)$ $\widehat{=}$ Menge der $k \in \mathbb{N}$ nächsten Nachbarn, bzgl. euklidischer Norm

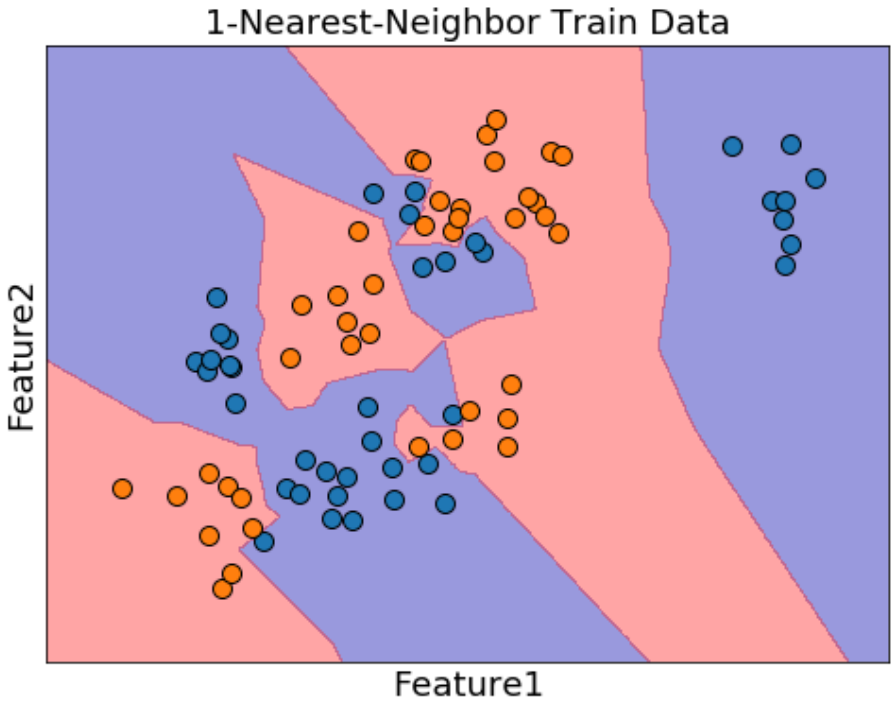
K-Nächste-Nachbarn



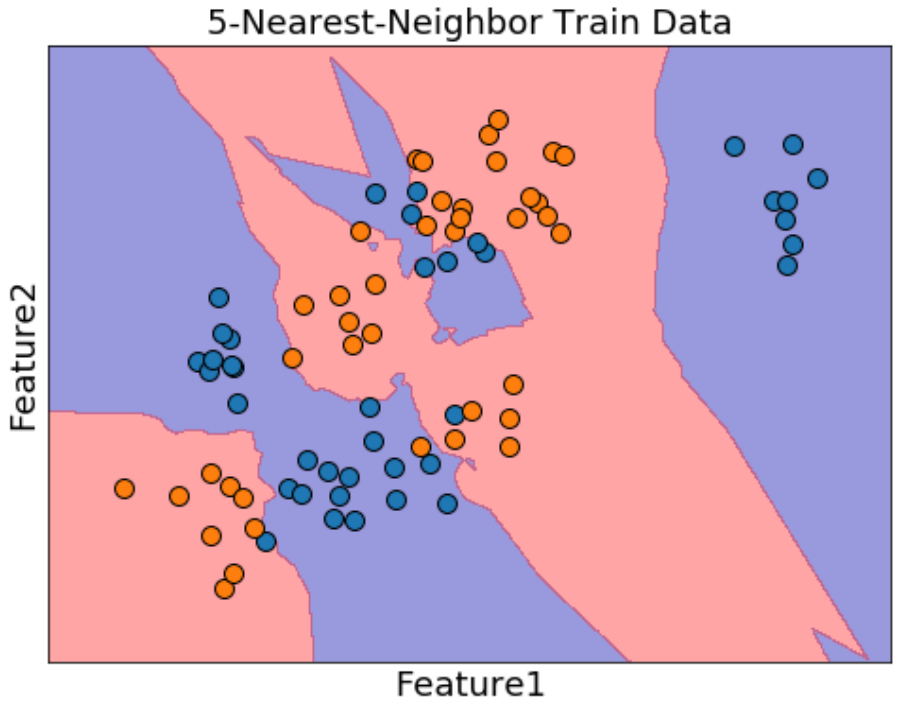


Universität zu Köln

K-Nächste-Nachbarn

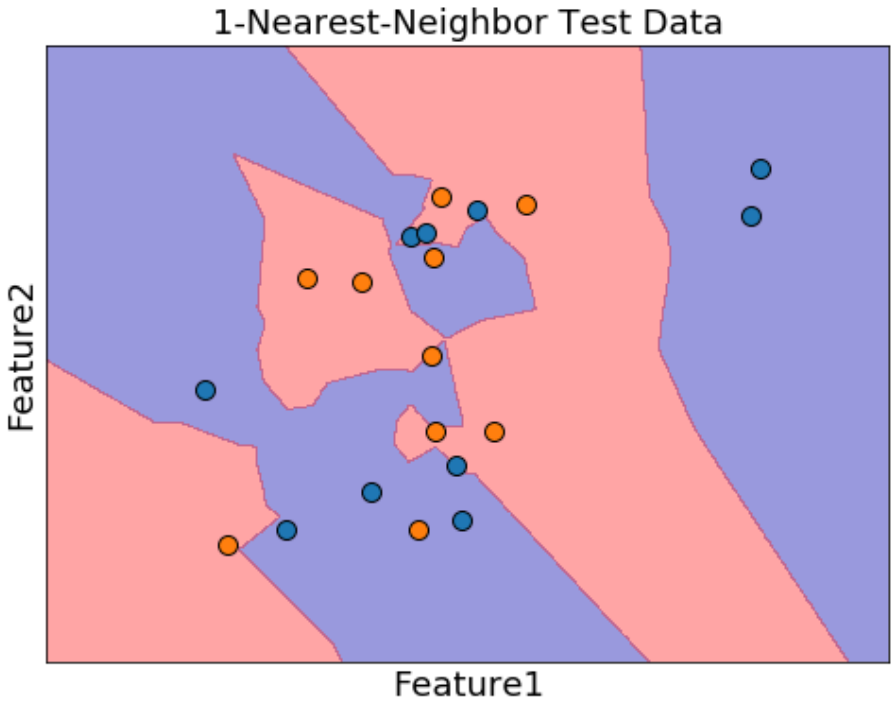


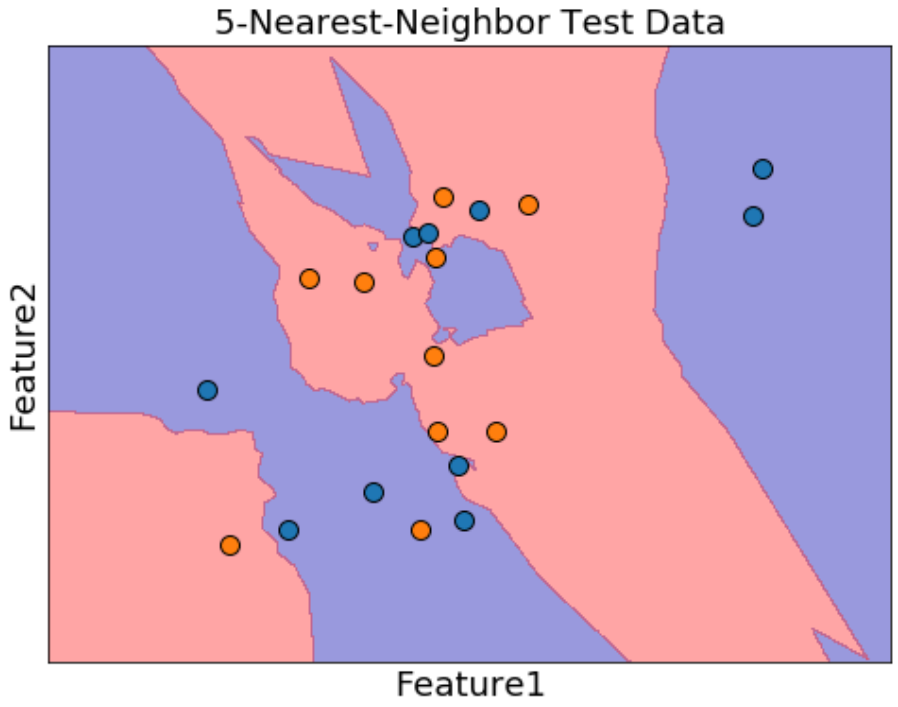
Genauigkeit: 100%



Genauigkeit: 94%

K-Nächste-Nachbarn - Testdaten

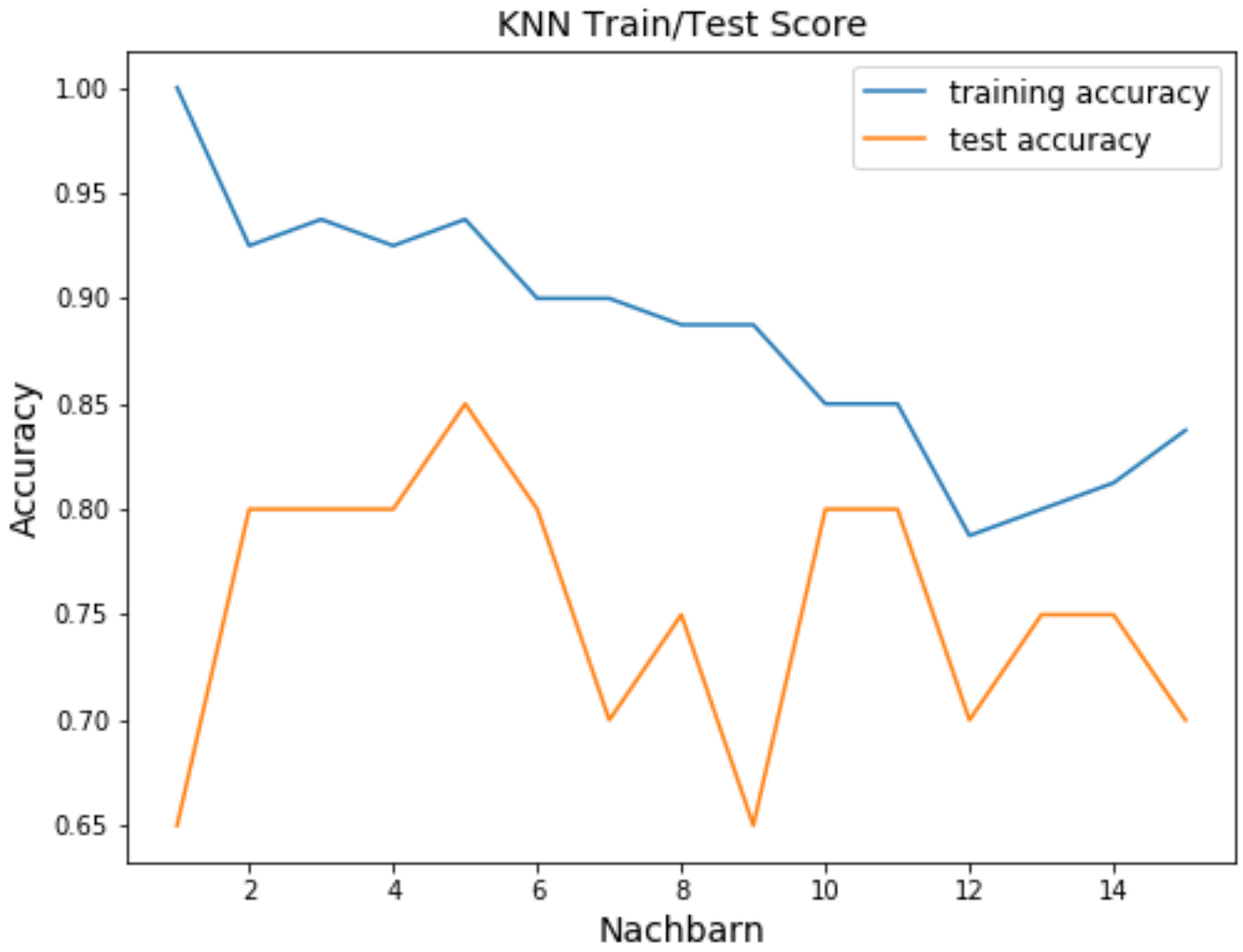




Genauigkeit: 65%

Genauigkeit: 85%

K-Nächste-Nachbarn – Train/Test Daten



- $k = 1$: zu stark an den Trainingsdaten angepasst
- $k = 12$: Modell nicht Komplex genug

Least Squares

- Annahme: Linearer Zusammenhang

$$Y = X\beta + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

$$\Rightarrow f(X) = E(Y|X) = E(X\beta + \varepsilon|X) = X\beta + E(\varepsilon|X) = X\beta$$

- Einsetzen in $EPE\big(f(X)\big)$ und Erwartungswert durch Mittelwert ersetzen:

$$EPE(f(X)) = E\big((Y - X\beta)^2\big) \approx \frac{1}{N}\sum_{i=1}^{N}(y_i - x_i^T\hat{\beta})^2$$

Least Squares

- Alternative Matrixschreibweise:

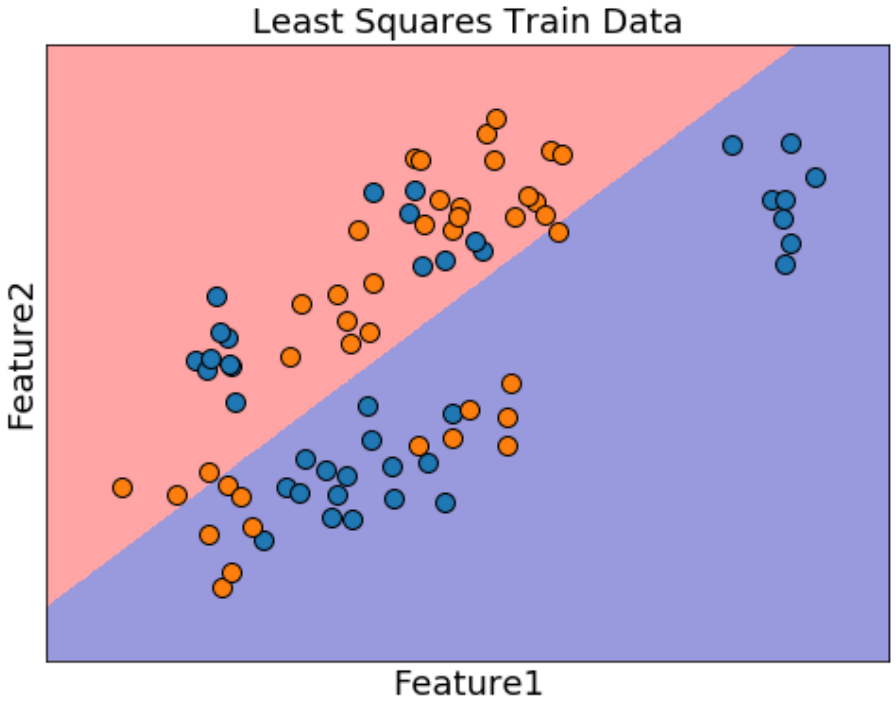
$$EPE(X) \approx \frac{1}{N}\sum_{i=1}^{N}(y_i - x_i^T\hat{\beta})^2 = \frac{1}{N}\left(y - X\hat{\beta}\right)^T(y - X\hat{\beta})$$

- Ableitung nach β gleich 0 setzen und umformen ergibt:

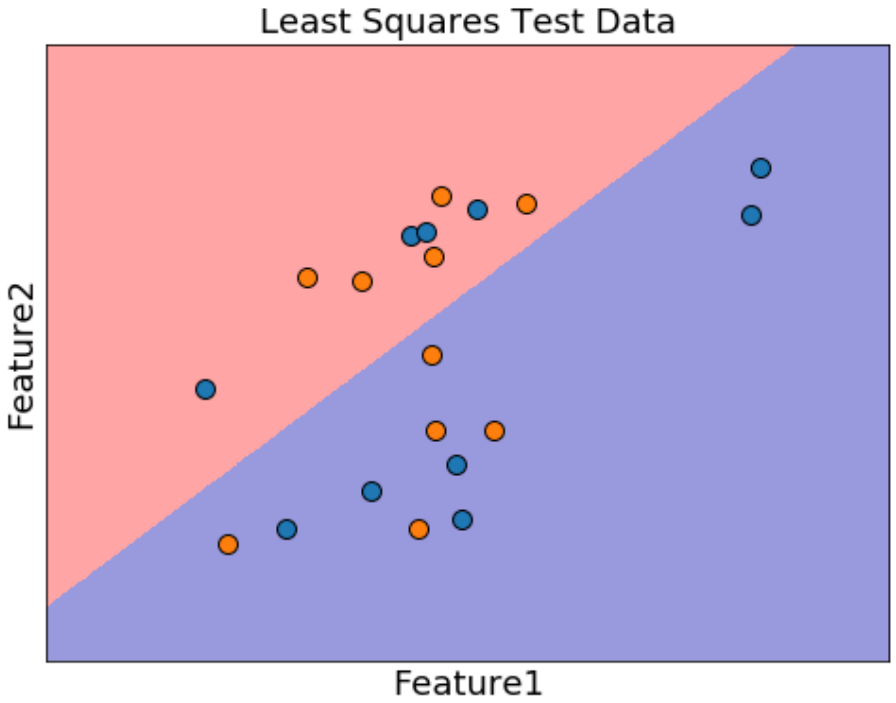
$$\frac{\partial EPE(X)}{\partial \hat{\beta}} = 0 \quad \Leftrightarrow \quad \hat{\beta} = (X^TX)^{-1}X^Ty$$

$$\Rightarrow \hat{y}_i = \begin{cases} 1 & falls \quad x_i^T\hat{\beta} \geq 0.5 \\ 0 & falls \quad x_i^T\hat{\beta} < 0.5 \end{cases}$$

Least Squares – Klassifikation

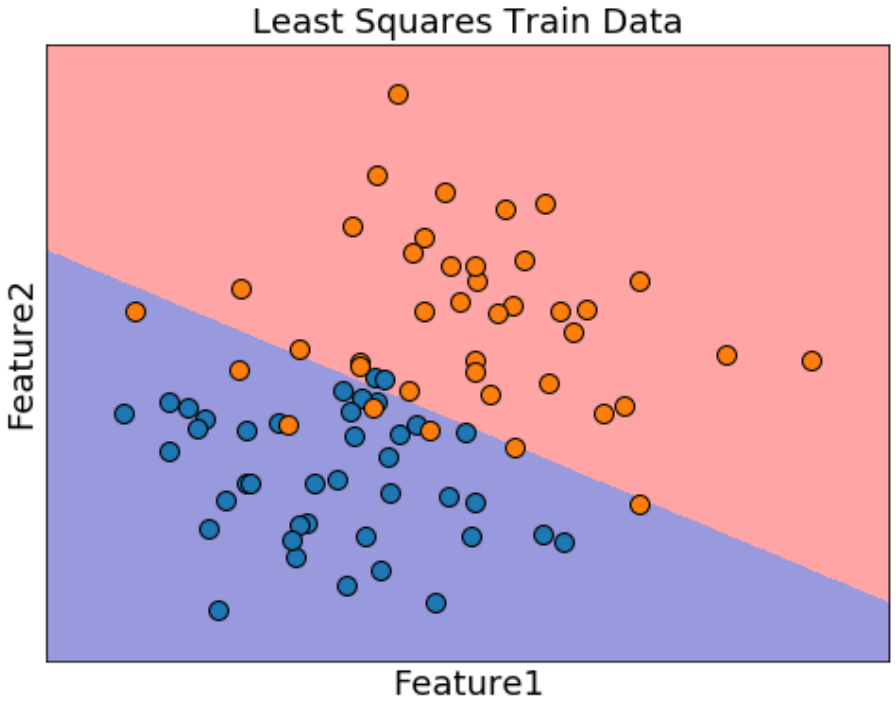


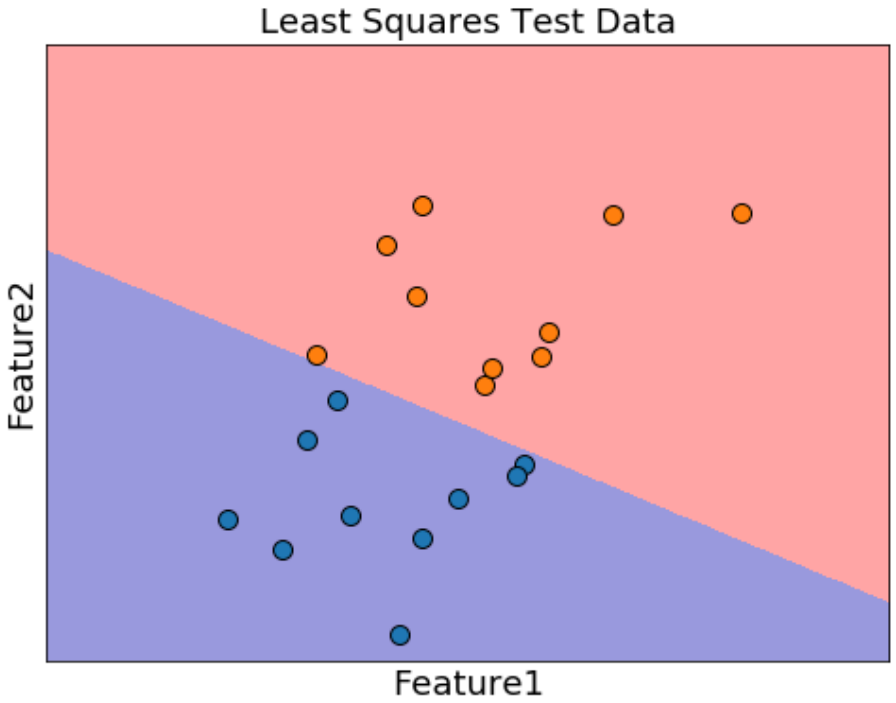
Genauigkeit: 63.75%



Genauigkeit: 55%

Least Squares – Bessere Verteilung

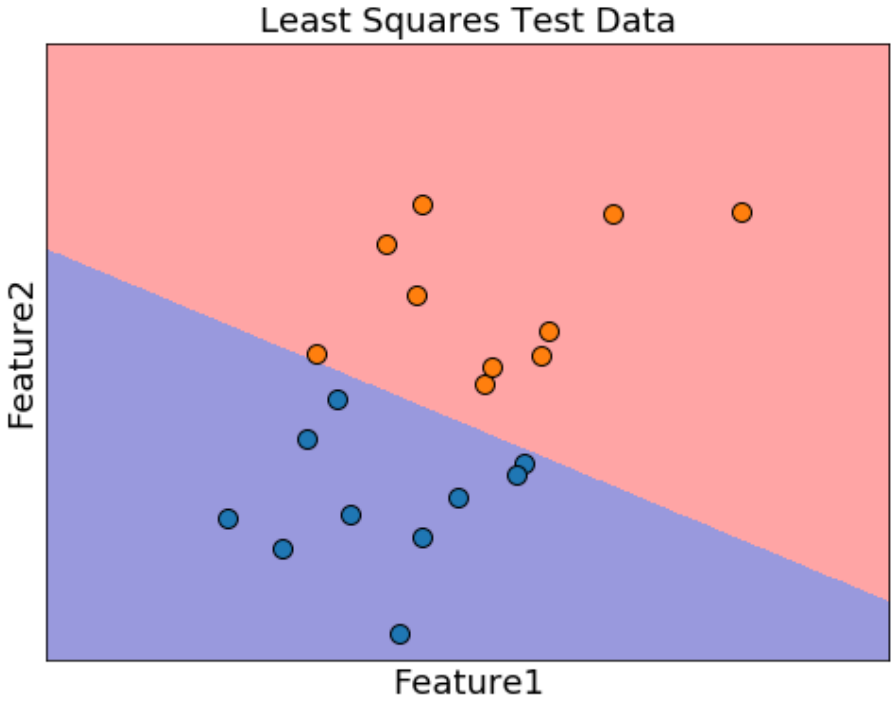


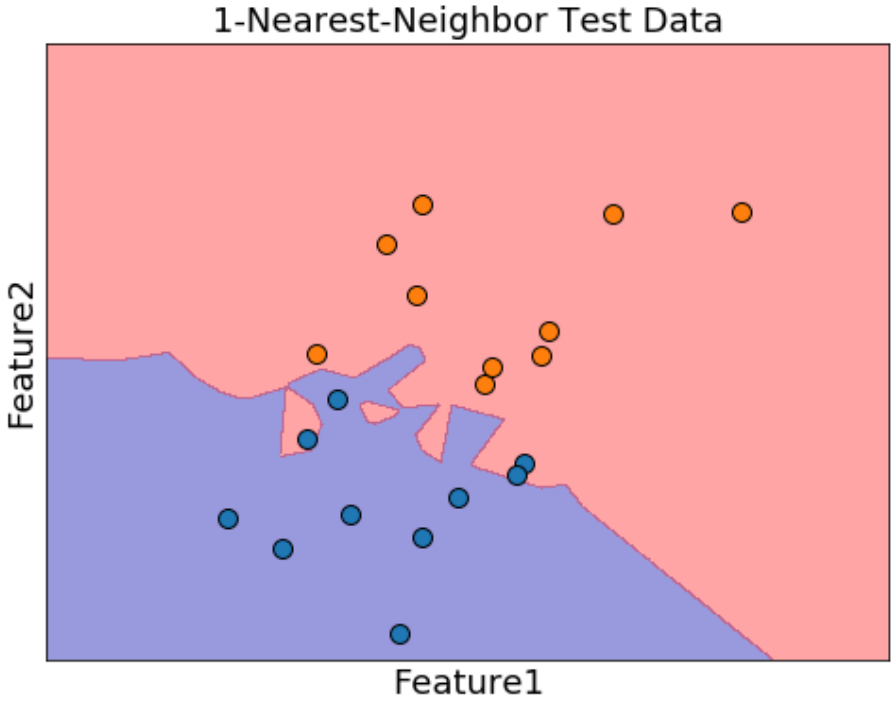


Genauigkeit: 88.75%

Genauigkeit: 100%

LS vs. KNN – Klassifikation



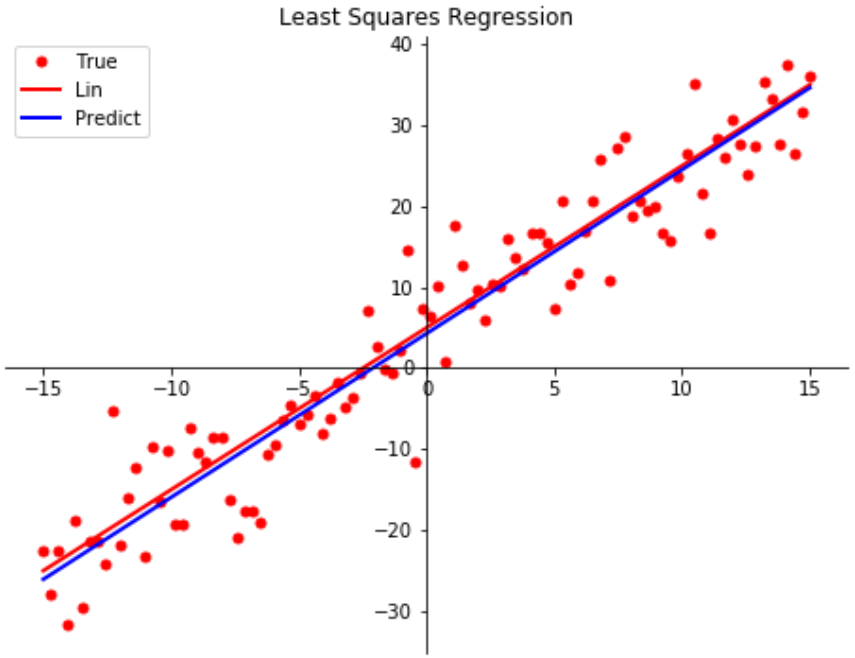


Genauigkeit: 100%

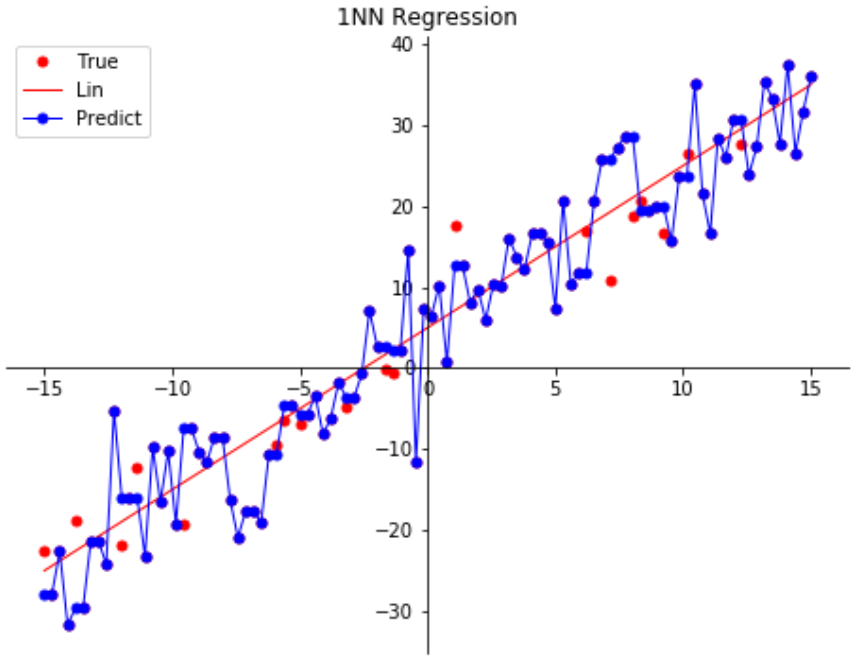
Genauigkeit: 85%

LS vs. KNN – Regression

$$y = 5 + 2x + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0,25)$$



Mean Squarred Error: 18.4

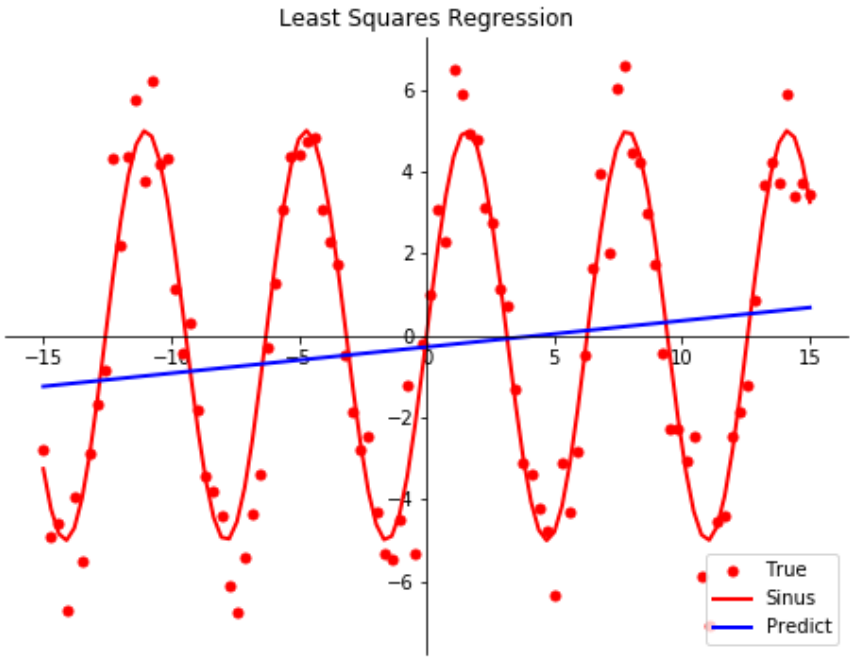


Mean Squarred Error: 37.7

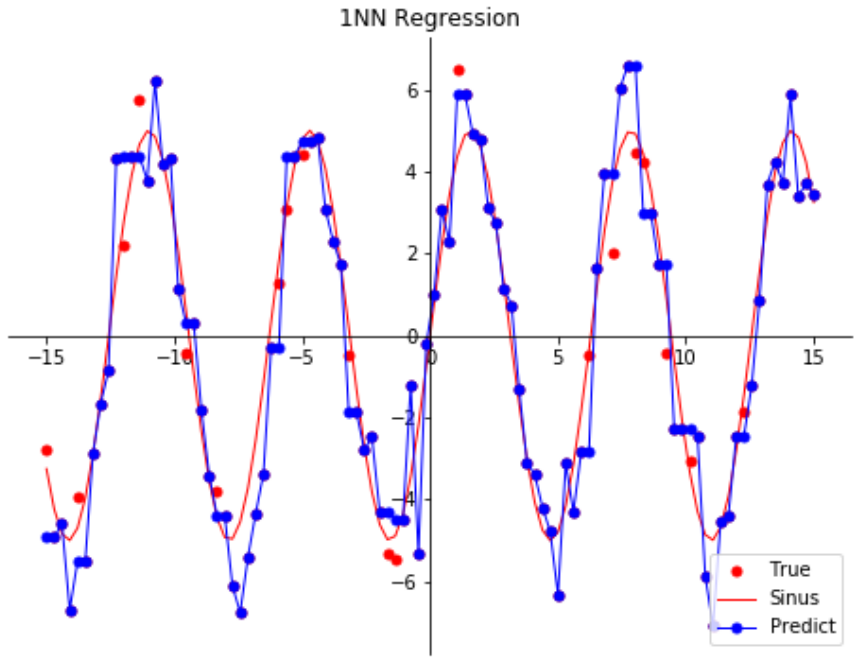
Universität zu Köln

LS vs. KNN – Regression

$$y = \sin(x) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0,1)$$



Mean Squarred Error: 13.4

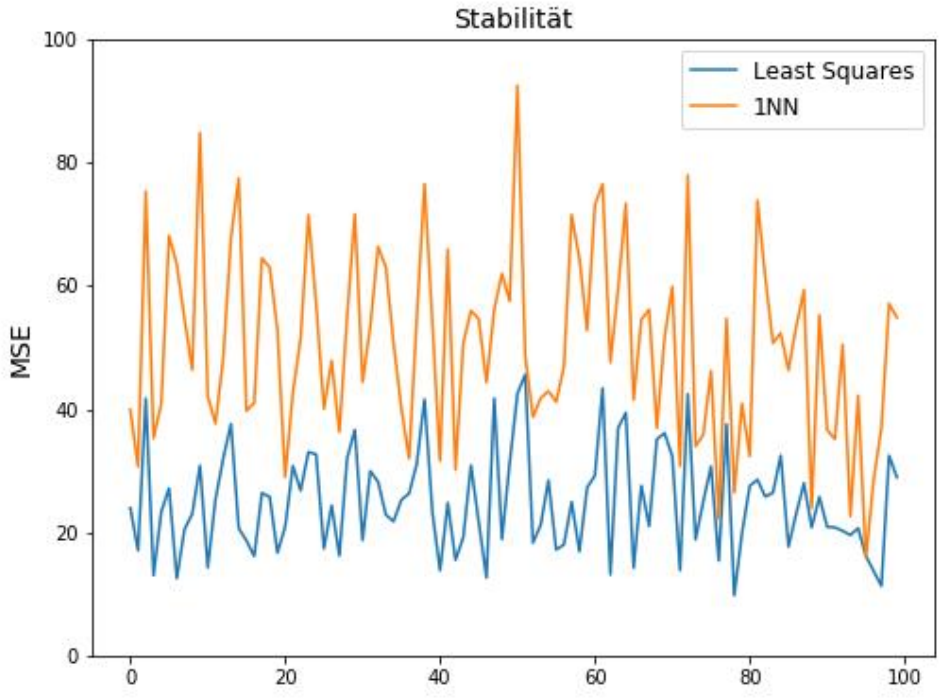


Mean Squarred Error: 2.2

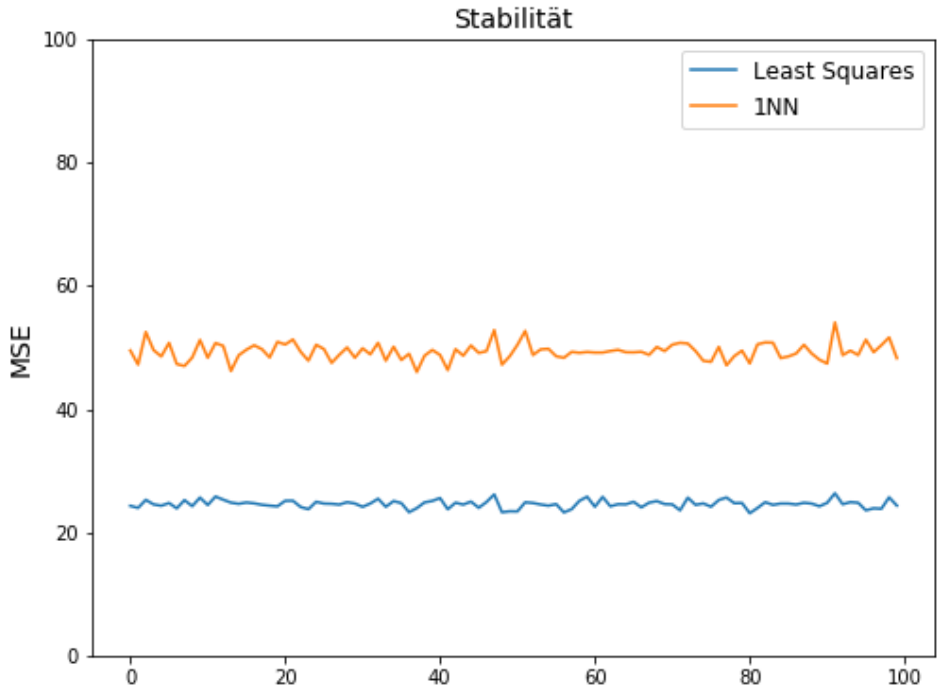
Universität zu Köln

Stabilität

$$y = 5 + 2x + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0,25)$$



$N = 100$



$N = 10000$

X-Achse: Unterschiedliche Aufteilungen der Daten in Trainings- und Testdaten

Bias-Varianz Dekomposition

- Mean Squared Error an der Stelle x_0 auswerten für:

$$y = f(x) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0,25)$$

$$MSE(x_0) = E_{\mathcal{T}}((y_0 - \hat{y}_0)^2)$$

$$= \left(f(x_0) - E_{\mathcal{T}}(\hat{y}_0)\right)^2 + (E_{\mathcal{T}}(\hat{y}_0^2) - E_{\mathcal{T}}(\hat{y}_0)^2) + \sigma^2$$

$$= Bias_{\mathcal{T}}^2(\hat{y}_0) + Var_{\mathcal{T}}(\hat{y}_0) + \sigma^2$$

Bias-Varianz Dekomposition

- $Bias_{\mathcal{T}}(\hat{y}_0)$: Misst wie gut das Modell den deterministischen Teil von y_0 approximiert

- $Var_{\mathcal{T}}(\hat{y}_0)$: Varianz der unterschiedlichen Trainings/Test Aufteilung. Wird kleiner je größer N ist.

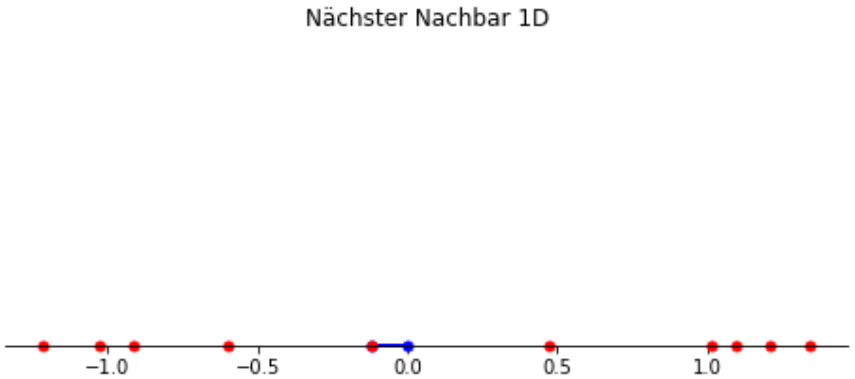
- σ^2: Varianz des Fehlerterms ε

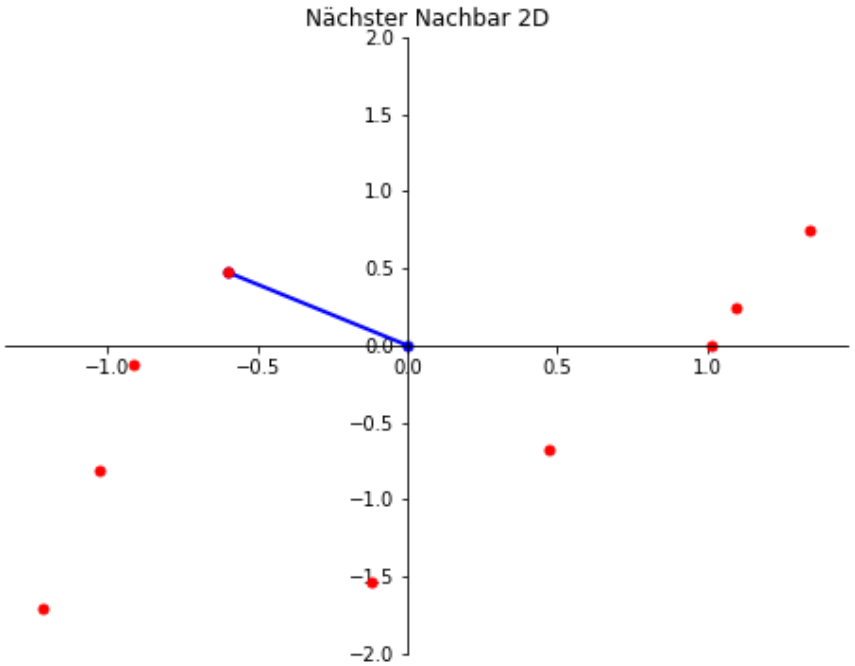
Bias-Varianz Dekomposition

- Least Squares:
  - $Var_{\mathcal{T}}(\hat{y}_0)$ wird klein für große N
  - $Bias(\hat{y}_0)$ ist klein, wenn alle Annahmen erfüllt sind, ansonsten ist auch Least Squares verzerrt.

- KNN:
  - $Var_{\mathcal{T}}(\hat{y}_0)$ wird klein für große N, jedoch größer als bei Least Squares
  - $Bias(\hat{y}_0)$ wird groß, falls der nächste Nachbar weit entfernt ist, besonders bei hohen Dimensionen

KNN 1D vs. 2D



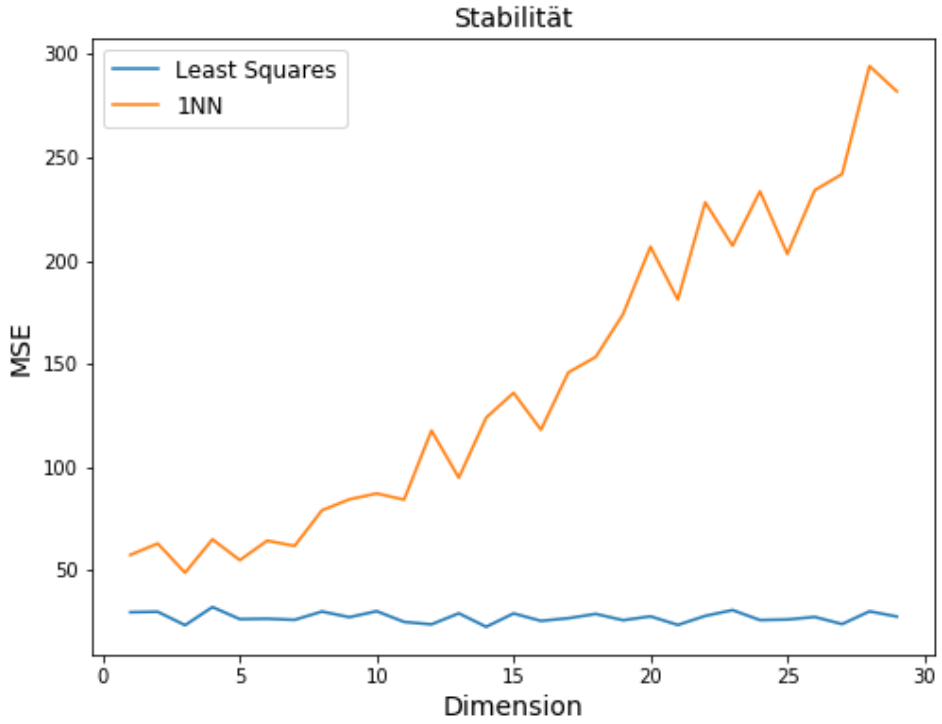


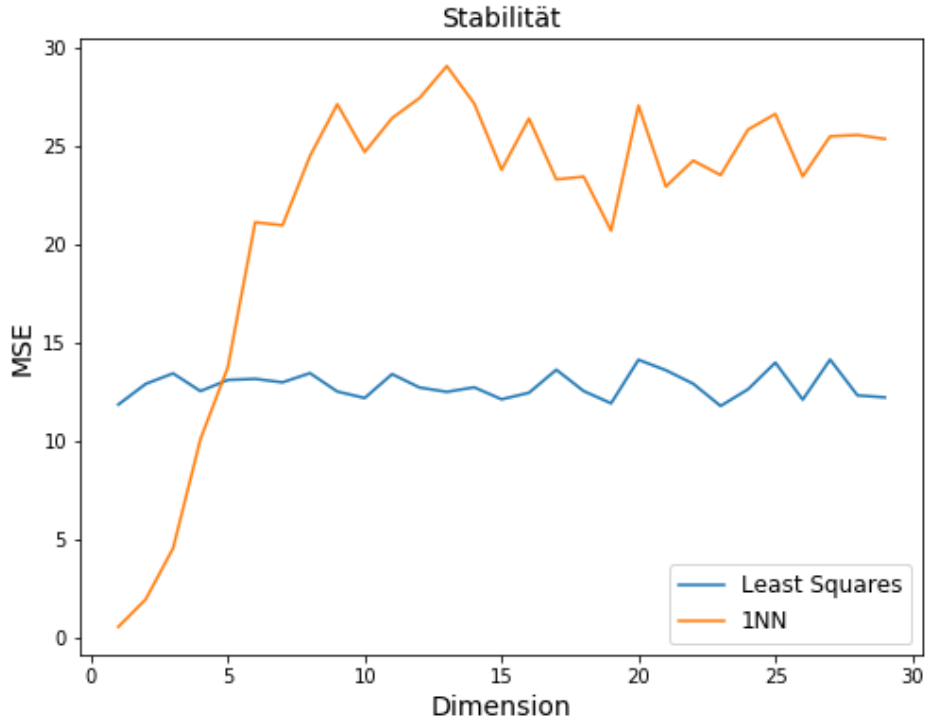
$$||x_0 - x_1|| = 0.12$$

$$||x_0 - x_1|| = 0.76$$

Universität zu Köln

LS vs. KNN – Hohe Dimensionen





$f(x) = 5 + \sum_{i=1}^{d} 2x_i$

$f(x) = 5\sin(\sum_{i=1}^{d} 2x_i)$

Universität zu Köln

Least Squares – Vor- und Nachteile

- Vorteile:
  - Stabil und unverzerrt
  - Gilt auch für höhere Dimensionen

- Nachteile:
  - Erfolg ist an oft unrealistischen Annahmen gebunden
  - Nur einfache Zusammenhänge können erklärt werden

KNN – Vor- und Nachteile

- Vorteile:
  - Keine Annahmen an die Daten notwendig
  - Komplexe Zusammenhänge können erfasst werden

- Nachteile:
  - Starke Verzerrung und Instabil
  - Vor allem für hohe Dimensionen

Fazit

- Wahl des Modells hängt stark von den zugrundeliegenden Daten ab
- Untersuchung der Daten ist ein wichtiger Bestanteil des Machine Learning Bereichs
- Anzahl der Beobachtungen sollte möglichst groß sein
- Anzahl der unterschiedlichen Features sollte möglichst reduziert werden

Vielen Dank für Ihre Aufmerksamkeit.